⑲日本国特許庁（JP）　⑪特許出願公開

⑫公開特許公報（A）　平2－105973

⑤Int. Cl.⁵	識別記号	庁内整理番号	㊸公開　平成2年(1990)4月18日
G 06 F 15/40	500 T	7313－5B	

審査請求　未請求　請求項の数 1　（全6頁）

㊹発明の名称　文書自動分類装置

㉑特　願　昭63－258748

㉒出　願　昭63(1988)10月14日

㉗発明者　河合　敦夫　東京都千代田区内幸町1丁目1番6号　日本電信電話株式会社内

㉗発明者　永田　昌明　東京都千代田区内幸町1丁目1番6号　日本電信電話株式会社内

㉗発明者　木本　晴夫　東京都千代田区内幸町1丁目1番6号　日本電信電話株式会社内

㉑出願人　日本電信電話株式会社　東京都千代田区内幸町1丁目1番6号

㉔代理人　弁理士　森田　寛

明細書

1. 発明の名称

文書自動分類装置

2. 特許請求の範囲

文書入力装置から入力された日本文文書データベースを取り扱う自然言語処理システムにおいて，

各分類分野ごとに偏って文書中に出現するキーワード及びキーワードの意味カテゴリを調べる分野特徴抽出装置と，

各分類分野ごとに偏って文書中に出現するキーワードを記述した分野識別単語得点表と，

各分類分野ごとに偏って文書中に出現する単語の意味カテゴリを記述した分野識別意味カテゴリ得点表と，

分類の対象となる文書中に出現したキーワードとその意味カテゴリとから分野識別単語得点表と分野識別意味カテゴリ得点表とを用いて文書の分類先を決定する分類先識別装置

を備えた

ことを特徴とする文書自動分類装置．

3. 発明の詳細な説明

(1) 発明の属する技術分野

本発明は，文書データベース作成のために，データベースに蓄積される文書に対して，その文書の分類を自動的に行う文書自動分類装置に関するものである．

(2) 従来の技術

新聞記事，特許出願書類，技術論文などの大量の文書を含むデータベースを作成する場合，データベースの入力の際に各文書に対して，分類用のコードを付与する必要が生じる．従来，この目的のために，ある分類分野に偏って出現する傾向の高い単語に着目する方法が用いられてきた．

この方法では，すでに分類済みの文書中の単語を統計的に処理して，各分野に偏って出現する単語（今後，分野識別単語と呼ぶ）を決定する．次

に，未分類の文書中の分野識別単語を手掛かりに，文書の分類先を決定する手法である。分野識別単語の例として，例えば，スポーツ，国際，…の分類分野を設定した時，単語"オリンピック"が分類分野"スポーツ"に，また単語"外交官"が分類分野"国際"に偏って出現する場合，"オリンピック"や"外交官"を分野識別単語とする。

しかし，分野識別の手掛かりとして，単語の表記（文字列）そのものを用いると，分野識別の点からは同一の単語集合として扱ってもよい，（i）表記のゆれ（例　コンピューターとコンピュータ，組合せと組み合わせなど），（ii）同義語（例　コンピューターと電子計算機，首相と総理大臣など），（iii）広い意味で同じ概念を表す単語集合（例　ゴルフ，剣道，フェンシング，…は，スポーツの概念を表す）が，別々の分野識別単語になる。

このため，未分類文書中に出現した単語が，分野識別単語として登録されている単語の1つと同じ概念を表していても，その分野識別単語との表記のずれがあったり，同義語である場合は，その分野識別単語と全く別の文字列として取り扱われる。このため，未分類文書中には分野識別単語が存在しないことになり，分類が不可能になる，という欠点があった。

(3) 発明の目的

本発明の目的は，単語の意味カテゴリを用いることにより，従来の文書自動分類装置の持つ上述の欠点を解決した文書自動分類装置を提供することにある。

(4) 発明の構成

(4-1) 発明の特徴と従来の技術との差異

従来の技術では，分野ごとに偏って出現する単語を分類装置に登録し，この単語を手掛かりに文書の分類を行っていた。本発明では，従来の技術に加えて，単語の意味カテゴリに着目し，分野ごとに偏って出現する意味カテゴリを，新たな手掛かりとして，文書の分類を行うことを，最も主要な特徴とする。

したがって，未分類文書中に分野ごとの特徴を表している分野識別単語と同じ概念を表すが文字列として異なる単語（表記のずれ，同義語）が表れた場合に，従来の技術では分類の手掛かりを得ることができなかったが，本発明では，意味カテゴリを用いることにより，同じ集合の単語として識別でき，分類の手掛かりが得られるという点で，従来の技術とは異なる。

(4-2) 実施例

第1図は，本発明をハードウェアによって構成した本発明の基本構成例を示す。1は分類コード付き文書ファイルで，分野ごとの文書の特徴を抽出するために用いる標本データである。2は分野特徴抽出装置で，分野ごとの文書の特徴を抽出する。分野ごとの文書の特徴（キーワードと意味カテゴリの出現頻度の分野ごとの偏り）を，それぞれ分野識別単語得点表3，分野識別意味カテゴリ得点表4へ記録する。5は分類コード無し文書ファイルである。6は分類先識別装置で，未分類の文書に分類コードを自動的に付与し，その結果を分類コードファイル7へ出力する。

次に，第2図を用いて，分野特徴抽出装置の説明を行う。まず，入力装置10から読み込まれた分類コード付き文書から，キーワード自動抽出・生成部11を介して，キーワード8が抽出または生成される。次に，キーワード頻度計算部12では，このキーワード8の出現頻度を，分類コードをもとに，分野別キーワード頻度表9へ加算する。意味カテゴリ検索部13では，日本語辞書15の意味カテゴリ記述部を検索して，キーワード8のそれぞれに意味カテゴリを付与する。次に，意味カテゴリ頻度計算部14でも同様に，キーワードの意味カテゴリ16の出現頻度を分類コードごとに，分野別意味カテゴリ頻度表17へと加算する。以上の操作を，分類コード付き文書の数だけ行う。

単語得点表計算部18では，不要キーワードの削除，頻度から得点への変換により分野識別単語得点表3を作成する。具体的には，分野別キーワー

ド頻度表9の中から，

① 全体としての出現回数が低いキーワード，

② 各分野にわたって，均一的に出現し，出現分野に偏りが少ないキーワード，

を頻度表から削除する。次に，各キーワードの各分野における頻度を得点へと変換する。意味カテゴリ得点表計算部19でも，分野別意味カテゴリ頻度表17をもとに，不要意味カテゴリの削除，頻度から得点への変換により分野識別意味カテゴリ得点表4を作成する。

次に，第3図を用いて，分類先識別装置の説明を行う。入力装置21より読み込まれた分類コード無し文書は，キーワード自動抽出・生成部22により，キーワード20が抽出・生成される。次に，キーワード得点加算部23では，キーワードの中から，分野識別単語得点表3に載っているキーワードの得点を分類分野ごとに加算する。意味カテゴリ検索部26では，日本語辞書28の意味カテゴリ記述部を検索して，各キーワードの意味カテゴリ29を検索する。次に，意味カテゴリ得点加算部27では，キーワードの意味カテゴリ29の中から，分野識別意味カテゴリ得点表4に載っている意味カテゴリの得点を分類分野ごとに加算する。分野判定部24では，キーワード得点加算部と意味カテゴリ得点加算部の得点を各分野ごとに単純加算し，一番得点の高い分野を文書の分類先として決定する。そして，出力装置25により，分類先を分類コードファイル7へ書き込む。

第4図は分野別キーワード頻度の一例を説明する図である。この例では，分類分野として，政治，経済，科学，…，スポーツ，国際の10分野を設定している。それぞれの分野の文書に，各キーワードが何回表れたかが示されている。キーワード"円相場"は，政治分野の記事に5回，経済分野の記事に56回出現している。第5図は，分野識別単語得点の一例を説明する図あり，第4図をもとに作成した。"全電通"，"オフサイド"は，全体としての出現頻度が，それぞれ，3回，1回と小さく，たとえ分野識別単語として登録しても，他の文書に出現する確率が低いので，第4図から削除する。また，"東京"，"新聞記事"は，各分類分野の文書に，平均的に出現するので，逆に，そのキーワードで分野を識別する手掛かりにはなりにくい。従って，分野識別単語としては不適切であり，第4図から削除する。

次に，こうして選択された分野識別単語jの，分野kの頻度Xjkを，この動作例では（式1）によって得点Yjkに変換する。こうして，第5図図示の得点を得る。

$$Yjk=(Xjk-Mjk)/Mjk \quad \cdots\cdots(式1)$$

Mjk：単語jのk分野における理論度数であり（式2）によって求める。

$$Mjk=\frac{1}{\text{分野総数}(=10)}\sum_{k=1}^{\text{分野総数}(=10)}Xjk \quad \cdots\cdots(式2)$$

第6図は意味カテゴリとキーワードの関係の一例を表す説明図である。例えば"参議院"，"郵政省"，"市役所"などは意味上から"行政機関"というキーワードにまとめられている。第7図は分野別意味カテゴリ頻度の一例を説明する図である。第8図は分野識別意味カテゴリ得点の一例を説明する図であり，不要意味カテゴリの削除，頻度から得点への変換により，第7図図示の頻度から作成した。第7図および第8図は第4図および第5図の場合と同様にして得られる。

第9図は，第3図分類先識別装置の一動作例を説明する図である。入力装置より読み込まれた入力文書30は，キーワード自動抽出・生成部により，キーワード31が抽出・生成される。図において意味カテゴリは（ ）で囲って示されている。次に，意味カテゴリ検索部により，各キーワードの意味カテゴリ32を検索する。次に，①キーワードの中から分野識別単語得点（第5図）に載っているキーワードの各得点を分類分野ごとに，加算する。しかし，ここでは，分野識別単語得点に載っているキーワードはないので，各分野の得点は0となる。②分野識別意味カテゴリ得点（第8図）に載っている意味カテゴリは〔スポーツ〕だけであるので，この（スポーツ）の得点を分類分野ごとに加算する。ここで，同じ意味カテゴリ，単語が複数回出現した場合は出現回数分を加算する。分野

別得点表３３における①，②の得点を各分野ごとに加算し，第９図図示の例では一番得点の高い分野“スポーツ”を文書の分類先として決定する。

(5) 発明の効果

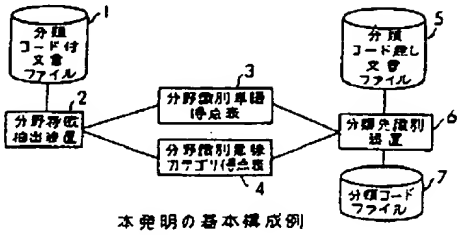
以上説明したように，本発明によれば，分類の手掛かりとして単語の意味カテゴリを用いるようにしているので，未分類文書中に，分野ごとの特徴を表している単語（分野識別単語）と同じ概念を持つが文字列としては異なる単語が出現した場合でも，同じ集合の単語として識別でき，分類の手掛かりが得られるという利点がある。

4. 図面の簡単な説明

第１図は本発明の基本構成例，第２図は分野特徴抽出装置の構成，第３図は分類先識別装置の構成，第４図ないし第８図は動作を説明するための説明図，第９図は動作例を示す図である。

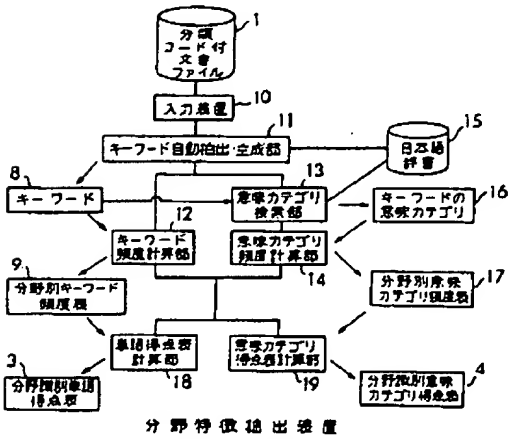
１…分類コード付き文書ファイル，
２…分野特徴抽出装置，
３…分野識別単語得点表，
４…分野識別意味カテゴリ得点表，
５…分類コード無し文書ファイル，
６…分類先識別装置，
７…分類コードファイル，
８…キーワード，
９…分野別キーワード頻度表，
10…入力装置，
11…キーワード自動抽出・生成部，
12…キーワード頻度計算部，
13…意味カテゴリ検索部，
14…意味カテゴリ頻度計算部，
15…日本語辞書，
16…キーワードの意味カテゴリ，
17…分野別意味カテゴリ頻度表，
18…単語得点表計算部，
19…意味カテゴリ得点表計算部，
20…キーワード，
21…入力装置，
22…キーワード自動抽出・生成部，
23…キーワード得点加算部，
24…分野判定部，
25…出力装置，
26…意味カテゴリ検索部，
27…意味カテゴリ得点加算部，
28…日本語辞書，
29…キーワードの意味カテゴリ，
30…入力文書，
31…キーワード，
32…意味カテゴリ，
33…分野別得点表。

特許出願人　日本電信電話株式会社
代理人　弁理士　森田　寛



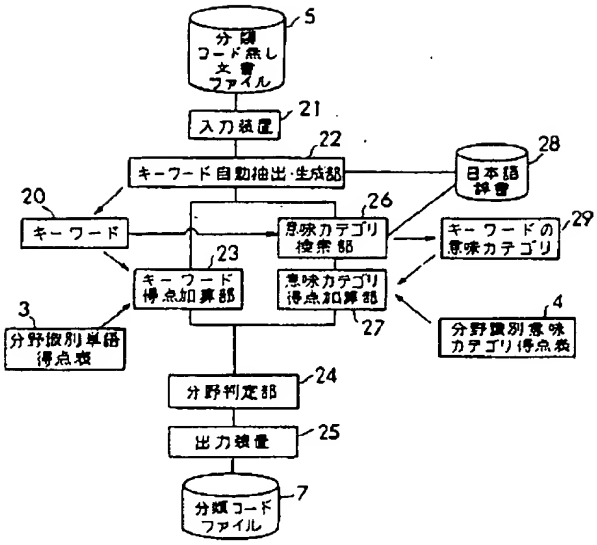
本発明の基本構成例

第　１　図



分野特徴抽出装置

第　２　図



分類先識別装置

第 3 図

キーワード ＼ 分野	政治	経済	科学	…	スポーツ	国際	全分野合計
円相場	5	56	0		0	3	70
＊東京	50	36	16		14	46	170
＊全電通	3	0	0		0	0	3
（キーワードj）				…Xjk			
…							
外交官	3	1	0		1	25	30
＊新聞記事	16	6	4		4	10	80
オリンピック	4	0	0		20	5	30
＊オフサイド	0	0	0		1	0	1

分野別キーワード頻度

第 4 図

識別単語 ＼ 分野	政治	経済	科学	…	スポーツ	国際
円相場	−0.29	7.0	−1.0		−1.0	−0.57
…				…Yjk		
…						
外交官	0.0	−0.67	−1.0		−0.67	7.33
オリンピック	0.33	−1.0	−1.0		5.67	0.67

分野識別単語得点

第 5 図

［行政機関］：参議院，郵政省，市役所，…
［専門的職業］：女優，教師，技師，画家，医者，…
［スポーツ］：ゴルフ，カヌー，選手権，プレー，競技会，…
［思想］：意向，意図，所存，念願，異存，…
［通信］：プロトコール，INS，…
［催し］：コンクール，個展，祭り，博覧会，展示会，…

意味カテゴリとキーワードの関係の一例

第 6 図

意味カテゴリ ＼ 分野	政治	経済	科学	…	スポーツ	国際	全分野合計
［行政機関］	120	36	5		2	3	200
＊［専門的職業］	50	36	36		24	6	270
［スポーツ］	3	0	1		90	1	100
（意味カテゴリj）				…Xjk			
…							
＊［思想］	50	26	4		4	20	210
［通信］	1	0	60		2	3	70
＊［催し］	30	10	10		10	20	200

分野別意味カテゴリ頻度

第 7 図

意味カテゴリ ＼ 分野	政治	経済	科学	…	スポーツ	国際
［行政機関］	5.0	0.8	−0.75		−0.9	−0.85
［スポーツ］	−0.7	−1.0	−0.9		8.0	−0.9
（意味カテゴリj）				…Yjk		
…						
［通信］	−0.86	−1.0	7.57		−0.71	−0.57

分野識別意味カテゴリ得点

第 8 図

入力文書 30

プレーオフは、米英感信かけ。
ゴルフの全米オープン選手権は、カーチス・ストレンジ（米）とニック・ファルド（英）が通算6アンダーで首位に並び、20日（日本時間21日）のプレーオフにもつれ込んだ。国の威信も賞金もかかっての一騎打ちとなるだけに、いってうの関心を集めそうだ。
ザ・カントリークラブでの同大会の開催は、3回目。前の2回もプレーオフにもち込まれたが、一騎打ちは今度が初めて。

キーワード 31	意味カテゴリ 32
プレー	[スポーツ，遊び]
オフ	[休業]
感信	[感（状態）]
ゴルフ	[スポーツ]
オープン	[開放，開始]
選手権	[競争，スポーツ，催し]
カーチス・ストレンジ	---未知語]
米	[国家]
ニック・ファルド	---未知語
英	[国家，言語]
通算	[計算]
アンダー	---未知語
日本	[国家]
時間	[時刻，授業]
国	[郷里，国家]
一騎打ち	[闘争]
関心	[興味]
ザ・カントリークラブ	---未知語
大会	[催し，会議]
開催	[活動]

分野 ＼	政治	経済	科学	・・・	スポーツ	国際 33
①	0.0	0.0	0.0	.	0.0	0.0
②	-2.1	-3.0	-2.7	.	24.0	-2.7
合計	-2.1	-3.0	-2.7	.	24.0	-2.7

分類先識別装置の一動作例

第 9 図